

Metadatenqualität und - interoperabilität

Jürgen Braun

16.03.2010



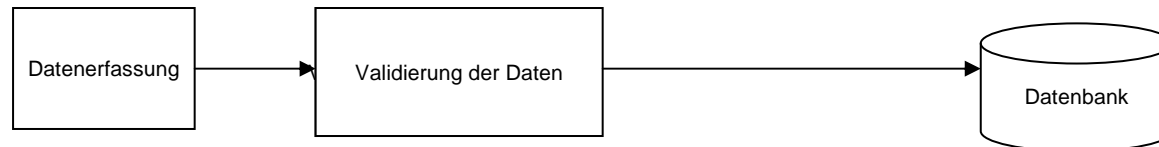
Interoperabilität von Metadaten

- **Probleme:**
 - Metadaten werden nicht nach einheitlichen Regeln und Standards erfasst
 - Die einzelnen Communities haben spezifische Anforderungen an ihre Metadaten, die von den vorhandenen Formaten oft nicht erfüllt werden
 - Selbst bei der Verwendung des gleichen Datenformats gibt es lokale Abweichungen und unterschiedliche Interpretationen von Katalogisierungsregeln
 - Fehlende oder unzureichende Dokumentation der Formate
 - Interoperabilität, d.h. der Austausch und die gemeinsame Nutzung von heterogenen Metadaten setzt voraus, daß die Metadaten nach einheitlichen Standards erstellt werden.
 - Die Qualität der Metadaten spielt eine wichtige Rolle bei der Interoperabilität



Anwendungsszenarien I

- Erfassung von Metadaten

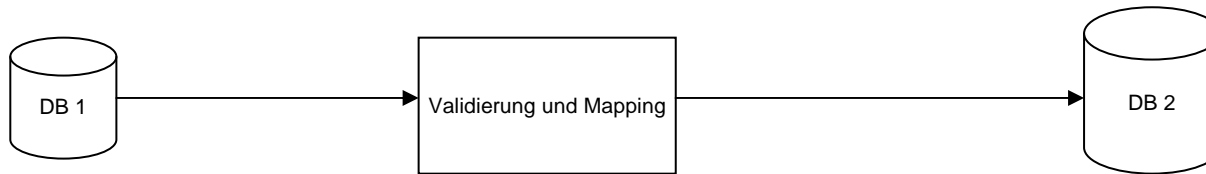


- Metadaten werden häufig von nichtbibliothekarischem Personal oder auch von den Verfassern selbst erstellt. Umfangreiche Regelwerkskenntnisse können daher nicht vorausgesetzt werden.



Anwendungsszenarien II

- Bereits vorhandene Metadaten sollen übernommen werden (Nachnutzung)



- Für das Mapping in andere Datenformate ist die syntaktische und semantische Korrektheit unverzichtbar um brauchbare Ergebnisse zu erzielen
- Ziel: Möglichst vollständige und semantisch korrekte Abbildung der Metadaten aus dem Quellformat ins Zielformat



Anwendungsszenarien III

- Verteilte Suche via SRU bzw. OAI



- Bei verteilten Suchanfragen mit SRU oder OAI müssen die Metadaten aus den verschiedenen Quellsystemen richtig interpretiert werden, um ein vollständiges und korrektes Suchergebnis zu erhalten
- Ziel: Syntaktisch und semantisch korrekte Abbildung der Metadaten aus verteilten Systemen



Kriterien zur Qualität von Metadaten

- Bruce und Hillmann (The continuum of metadata quality, 2004) haben sieben Kriterien zur Bestimmung der Metadatenqualität definiert:
- **Completeness**
 - sind alle für die Ressource relevanten Elemente vorhanden?
- **Accuracy**
 - Rechtschreibung, Abkürzungen, Namensansetzungen ...
- **Provenance**
 - läßt Rückschlüsse auf die Metadatenqualität zu
- **Conformance to expectations**
 - Erwartungen der Zielgruppe (in Bezug auf Metadatenelemente, kontrolliertes Vokabular)
- **Logical consistency and coherence**
 - Verwendung von einheitlichen Standards und Vokabular



Kriterien zur Qualität von Metadaten

- **timeliness**
 - aktuelle Synchronisierung der Metadaten, mit der Ressource die beschrieben wird
 - Zeit zwischen der Veröffentlichung der Ressource und der Erstellung der Metadaten
 - Metadaten in Bibliothekskatalogen werden nur selten aktualisiert, bei elektronischen Ressourcen ist das problematisch
- **Accessibility**
 - Metadaten müssen für den Nutzer sichtbar und verständlich sein (Format, techn. Voraussetzungen)
 - Berücksichtigung der Bedürfnisse der Zielgruppen



Qualität von Metadaten

- **Qualität von Metadaten wird bestimmt durch**
 - Syntaktische Korrektheit der Metadaten
 1. Wohlgeformtheit (maschinell)
 2. Übereinstimmung mit dem Metadatenprofil (maschinell)
 - Semantische Übereinstimmung mit den im Metadatenprofil festgelegten Definitionen (maschinell nur begrenzt möglich)
 - Entsprechen die Daten nicht der im Metadatenprofil definierten Syntax und Semantik sind Probleme bei der Übernahme der Daten in andere Systeme vorprogrammiert.



Möglichkeiten zur Sicherung der Qualität von Metadaten

- Die Qualität der Metadaten sollte idealerweise schon bei der Erfassung der Daten durch entsprechende Validierungsroutinen sichergestellt werden.
- Dokumentation des Metadatenprofils für Menschen und Maschinen:
 - Beschreibung des Profils, Userguides, Best Practise
 - RDF, XML Schema
- Maschinelle Validierung von bereits vorliegenden Metadaten durch Software-Tools
 - Aber: nicht alle Probleme werden von den Software-Tools erkannt
- Manuelle Validierung der Syntax und Semantik von bereits vorhandenen Metadaten
 - Die zeitaufwendige manuelle Validierung kann durch Software-Tools erleichtert werden



Kriterien für die Validierung

- **Syntaktische Validierung**
 - Sind alle verwendeten Datenelemente im Metadatenprofil definiert?
 - Sind alle Pflichtfelder vorhanden?
 - Wiederholbare Elemente
 - Werden nur die zugelassenen Datentypen verwendet (z.B. beim Datum)?
 - Ist die Zeichenkodierung korrekt?
- **Semantische Validierung**
 - Kontrolliertes Vokabular anhand der Encoding Schemes überprüfen
 - Abhängigkeiten zwischen Datenelementen überprüfen (z.B. wenn ein Encoding Scheme angegeben ist, muss auch ein gültiger Wert vorhanden sein)
 - Inhaltliche Konsistenz von nichtkontrolliertem Vokabular



Anforderungen an ein Validierungs-Werkzeug

- Ein Abgleich der Syntax mit dem Metadaten-Profil muss automatisiert möglich sein
- Software muss für unterschiedliche Metadatenprofile einsetzbar sein (generisches Tool)
- Unterstützung von verschiedenen Formaten (XML, Daten in Tabellenform oder ISO-Dateien)
- Unterstützung der wichtigsten Zeichensätze (UNICODE, UTF-8, ISO)
- Software sollte erweiterbar und flexibel sein (Anpassung an neue Metadatenformate)
- Verwendung von offenen Standards (XML, SOAP, WSDL)
- Software muss in andere Anwendungen integrierbar sein



Vascoda Checker Tool

- Wurde zur Validierung von Vascoda AP Metadaten an der SUB Göttingen entwickelt
- Vorteile:
 - Web-Anwendung mit graphischer Oberfläche
 - Einfache Bedienbarkeit
 - Statistische und grafische Auswertung der Metadatenelemente
 - Übersichtliche Darstellung der Ergebnisse
 - Deutschsprachige Dokumentation
- Nachteile
 - Kein generisches Tool, zur Validierung von anderen Metadatenformaten muss der Quellcode verändert werden
 - Nicht in andere Anwendungen integrierbar



Vascoda Fehler-Report

vascoda checker-tool

START DOKUMENTATION QUALITÄT STATISTIK BERICHT

OLC-SSG Astronomie senden

Modul: OLC-SSG Astronomie

Beim Klick auf die fett gedruckten Fehler und Warnungen gelangen Sie zur Anzeige des gewählten Elements (Anzeige "Statistik").

Allgemein		
Anzahl an Datensätze		8.229
richtige Datensätze	0%	0
fehlerhafte Datensätze	100%	8.229

Verpflichtende Elemente (Fehler)		
fehlerhafte Datensätze, bei denen ein oder mehrere verpflichtende Elemente fehlen	100%	8.229
dc:subject oder vap:thematic oder vap:ddc oder dcterms:spatial oder dcterms:temporal fehlen		8.229
dc:title fehlt		0
dcterms:issued fehlt		0
dc:type fehlt		0
dc:identifier und dcterms:bibliographicCitation fehlen		0
vap:resourceURI fehlt		0

Wiederholende Elemente (Fehler)		
Datensätze, in denen eine Wiederholung vorkommt, die nicht erlaubt ist (extent, issued, medium, edition)	0%	0

Sonstige Fehler		
dc:language liegt nicht im Richtigen Format vor		0
dcterms:issued liegt nicht im richtigen Format vor		0

Fertig



Vascoda: Statistische Auswertung

vascoda checker - Mozilla Firefox

http://arbeit.hilses.de/analyse.php?navi=Statistik

vascoda checker tool

helpdesk | logout | impressum
Sie sind eingeloggt als: vascoda Intern

START DOKUMENTATION QUALITÄT STATISTIK BERICHT

OLC-SSG Astronomie senden

Modul: OLC-SSG Astronomie

Es sind insgesamt 8.229 Datensätze vorhanden.

Beim Klick auf die fett gedruckten Fehler und Warnungen öffnet sich jeweils ein Fenster, in dem die fehlerhaften Einträge im entsprechenden Element mit dem Link zur Ansicht des Datensatzes ausgegeben werden. Das Fragezeichen gibt weitere Hinweise zu dem jeweiligen Element.

dc:title	
Element wird in 8.229 Datensätzen verwendet	
Fehler: 0	
Warnungen: 1	
korrekte Einträge: 8.228	

dc:creator	
Element wird in 7.522 Datensätzen verwendet	
Fehler: 0	
Warnungen: 0	
korrekte Einträge: 28.884	

dc:publisher	
Element wird in 8.005 Datensätzen verwendet	
Fehler: 0	
Warnungen: 0	
korrekte Einträge: 8.005	

Fertig



Vascoda Fehlermeldungen





Falcon

- Software zur Konvertierung und Validierung von Metadaten
- Basiert auf der Allegro Export-Sprache
- Vorteile:
 - Generisches Tool, im Prinzip können alle Metadatenformate eingelesen werden
 - Komplexe Validierungsroutinen können parametrisiert werden
- Nachteile
 - Kein Open-Source Projekt
 - Dokumentation unvollständig
 - Für komplexere Validierungen relativ hoher Einarbeitungsaufwand
 - XML wird nicht unterstützt
 - Keine Web-Anwendung
 - Windows-Anwendung, keine Plattformunabhängigkeit



Falcon: Anzeige eines Datensatzes

```
Falcon 5.0 - [F:\USER\EROMM\Prod\Partner\Paris\BNF_DC_0911.iso]
File Format Edit Index Search View Export Window Tools Help
000 01079nam 22002773n 450
001 FRBNF300000560000009:NUMM-5477721
009 http://catalogue.bnf.fr/ark:/12148/cb3000000560
039 $oCRI$aCG000100050401P
100 $a19970701d1856 m yOfrey50 ba
101 0 $afre
102 $aBE
105 $a||||z 00||||
106 $ar
200 1 $aAnathèmes et louanges. Les Régions du ciel, par Auguste Abadie
$bTexte imprimé
210 $aBruxelles$cJ.-B. Tarride$d1856
215 $aIn-16
686 $a240$2Cadre de classement de la Bibliographie nationale française
700 |$310340771$aAbadie$bAuguste$4070
801 0$aFR$bBNF$c19970701$gAFNOR$2intermrc
856 4 $uhttp://catalogue.bnf.fr/ark:/12148/bpt6k5477721n
913 2 $adumn mmmuu&
921 $aParis$cBibliothèque nationale de France$d2008$pDocument libre de
droits
922 $aFR$b2380
923 $a1 fichier (1297 octets)$cplusieurs formats (TIFF et JPEG)
951 $aFR$bBibliothèque nationale de France$dYE-13697$fa
952 $aFR$bBibliothèque nationale de France$cNUM$dNUMM-5477721
i #700 $a
100% - no more occurrences #1 : 1079 L:1 C:1
```



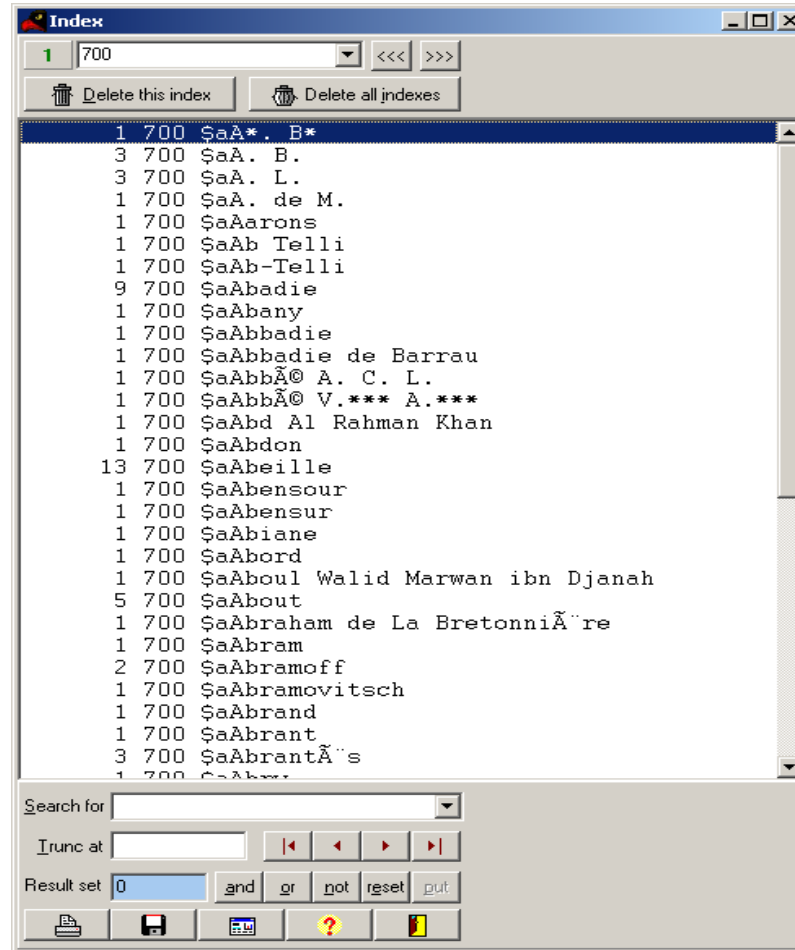
Falcon: Statistische Auswertung

tag/sub	count	occ max
200	56847	2
ind-1		
0	161	
1	56686	
ind-2		
-	56698	
1	149	
\$a	57054	6
\$b	56849	2
\$c	162	7
\$e	7883	3
\$f	8657	5
\$g	1198	3
\$h	182	2
\$i	362	3
205	1060	2
ind-1		
-	1060	
ind-2		
-	1060	
\$a	1058	1
\$b	7	1
\$f	61	1

More statistics reports are available via menu "File" or toolbar



Falcon: Index von Namen





Falcon: Validation Report

field	count	message
009	56847	invalid tag
010 \$b	3	subfield is not repeatable
010 \$d	2	subfield is not repeatable
011	468	indicator 1 invalid
039	62403	invalid tag
101	1	missing mandatory field
101 \$a	189	invalid language code
101 \$g	2	invalid language code
102 \$a	275	invalid country code
106	1	invalid length: 12
106 \$a	1	subfield is not repeatable
200	1	tag is not repeatable
200	136	indicator 2 invalid
205 \$a	2	missing mandatory subfield
210	167	tag is not repeatable
210	169	indicator 1 invalid
225	73	indicator 2 invalid
302 \$a	3	subfield is not repeatable
304 \$6	1	invalid subfield
304 \$7	2	invalid subfield
321 \$c	10	invalid subfield
326 \$b	345	missing mandatory subfield
423	3	subfield(s) invalid/not re:
503 \$d	25	subfield is not repeatable
503 \$h	47	subfield is not repeatable
503 \$j	44	subfield is not repeatable
510	260	indicator 1 invalid
510 \$6	2	invalid subfield
510 \$7	2	invalid subfield



Validierung von XML Daten

- **Grammatik-basierte Validierung anhand von:**
 - DTD
 - XML-Schema: Empfehlung des W3C, unterstützt unterschiedl. Datentypen und Namespaces
 - Für die wichtigsten bibliothekarischen Metadatenformate liegen Schema Definitionen vor z.B. MARCXML, MODS.
 - Für XML-Schema gibt es diverse frei verfügbare Validatoren z.B.
- **Xerces-C++**
 - Open source XML Parser, validiert Metadaten gegen DTD oder Schema
 - Nachteile: Zum Teil nur schwer verständliche Fehlermeldungen
 - Nur Validierung der Syntax möglich



Regel-basierte Validierung von XML Daten

- **Schematron**
 - Sprache zur Validierung von Metadaten
 - Ermöglicht eine regel-basierte Validierung
 - Wie Fehlermeldungen ausgegeben werden, kann durch ein Stylesheet gesteuert werden z.b. in HTML
 - Verwendet XSLT und XPATH Ausdrücke zur Validierung
 - Assertions (Behauptung/Annahme) werden zu einem Schema zusammengefasst
 - Falls eine Assertion unwahr ist kann eine Reaktion ausgeführt werden in der Regel eine Fehlermeldung.
 - Ermöglicht die Überprüfung von Regeln wie z. B.: Wenn ein Element X das Attribut A hat muss Element Y das Attribut B haben.



Validierung mit Schematron

- **Vorteile:**
 - sehr flexibel durch die Verwendung von XPath und XSLT Sprachelementen
 - erlaubt die Formulierung von eigenen Fehlermeldungen
 - Schematron Style-Sheets können sehr einfach in komplexe Anwendungen eingebunden werden
 - Zur Ausführung von Schematron ist nur ein XSLT Prozessor nötig z. B. Saxon
 - Graphische Benutzeroberflächen sind ebenfalls verfügbar z. B. von Topologi
 - XSLT und XPATH sind gut dokumentiert
- **Nachteile**
 - Es ist nur die Evaluierung von XML Metadaten möglich

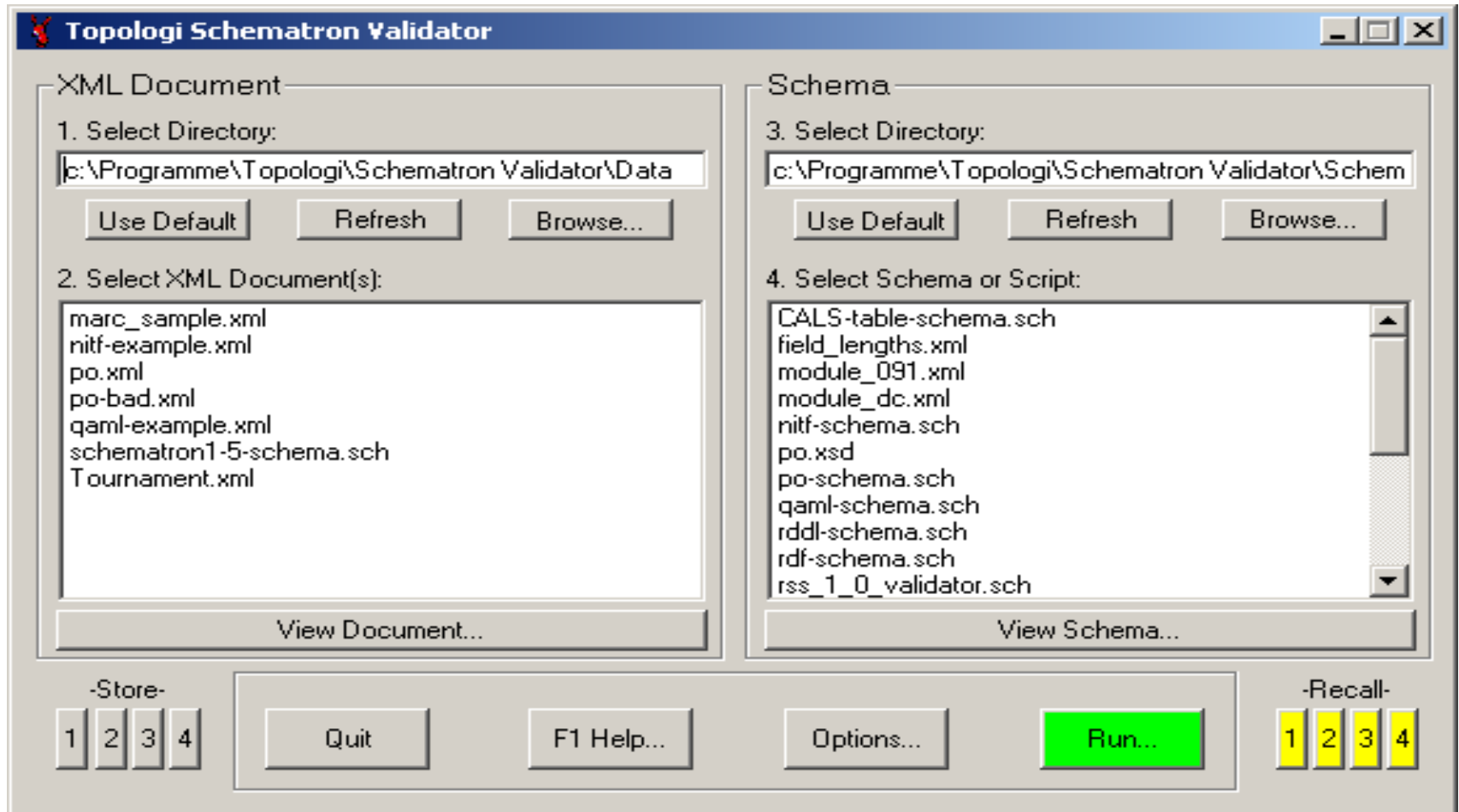


Schematron als Validierungs-Werkzeug

- Von den untersuchten Tools erfüllt nur Schematron die am Anfang formulierten Anforderungen (mit der Einschränkung, dass nur XML Daten damit validiert werden können)
- Mit Schematron kann die Validierung für unterschiedliche Metadatenprofile in XML für die eingangs beschriebenen Anwendungsszenarien weitgehend automatisiert werden
- Schematron Stylesheets sind kompatibel zu anderen offenen Standards (SOAP, WSDL) und plattformunabhängig
- Zur Erstellung von Schematron Stylesheets sind nur Kenntnisse in XPATH und XSLT nötig



Schematron Validator







Schematron Report

Validation Results - Browser Transformation - po-bad.xml (1 of 1)


Schematron Report

Schema for Purchase Order Example

Purchase Orders

-  A purchase order should have a billTo element.
`/purchaseOrder[1]`
`<purchaseOrder>...</>`
-  A purchase order should have an orderDate attribute.
`/purchaseOrder[1]`
`<purchaseOrder>...</>`

US Address

-  An address should have a state.
`/purchaseOrder[1]/shipTo[1]`
`<shipTo place="AU">...</>`

```
<!-- Deliberate errors! --><!-- See http://www.w3.org/TR/xmlschema-0/ --><purchaseOrder>
  <shipTo place="AU">
    <name>Kippabboo Versace</name>
    <street>18 Action Street</street>
    <city>Sleepy Town</city>
    <zip>2099</zip>
  </shipTo>
  <billAt country="NZ">
    <name>Katapo Feelgood</name>
    <street>1 Xena Avenue</street>
    <city>Auckland</city>
    <state>North Island</state>
    <zip>123</zip>
  </billAt>
  <comment>Hurry, my kiwi is going <i>wild</i>!</comment>
  <items>
    <Item partNum="872-AA">
      <ProductName>Deluxe Kiwi Beakbrush</ProductName>
      <Quantity>1</Quantity>
      <Price>148.95</Price>
      <Comment>Confirm this is electric</Comment>
    </Item>
  </items>
</purchaseOrder>
```

First Previous Next Last

Save Current... Repeat All Print Current... Close



Grenzen der maschinellen Validierung

- Eine semantische Evaluierung kann nur für kontrolliertes Vokabular automatisiert werden
- Bei nicht kontrolliertem Vokabular (z.B. Titel einer Ressource) kann nicht überprüft werden, ob der Inhalt der Elemente semantisch korrekt ist.
- Für die semantische Evaluierung von nicht-kontrolliertem Vokabular ist deshalb zur Zeit noch eine Überprüfung durch Fachpersonal nötig



Manuelle Validierung durch Fachpersonal

- Die semantische Validierung durch Fachpersonal ist sehr arbeitsintensiv, da Stichproben in der Regel nicht ausreichen, um alle Probleme zu erkennen
- Optimierung der manuellen Validierung durch geeignete Software-Tools:
 - Indexerstellung zur Kontrolle von einzelnen Datenelementen (Inkonsistenzen finden sich häufig am Anfang und Ende alphabetischer Indices)
 - Graphische Aufbereitung für den schnellen Überblick über große Datensammlungen. Probleme können so schneller erkannt werden
 - Die NSDL verwendet zum Beispiel Spotfire, eine kommerzielle Software zur graphischen Auswertung von grossen Datenmengen



Fazit und Ausblick

- Zum jetzigen Zeitpunkt erscheint die maschinelle Validierung allein noch nicht ausreichend um die Qualität von heterogenen Metadaten sicherzustellen.
- Aber: Durch den konsequenten Einsatz von gut dokumentierten (RDF, DCAM, Userguides) Anwendungsprofilen und kontrolliertem Vokabular (SKOS, OWL) lassen sich sowohl die Qualität bei der Erfassung von Metadaten, als auch die Möglichkeiten der maschinellen Validierung deutlich verbessern.



Metadatenqualität und -interoperabilität

- Vielen Dank für Ihre Aufmerksamkeit